

METHOD AND APPARATUS FOR PRODUCING NATURAL SOUNDING PITCH CONTOURS IN A SPEECH SYNTHESIZER

Field of the Invention

5 The present invention relates generally to speech synthesis systems and, more particularly, to methods and apparatus that generate natural sounding speech.

Background of the Invention

10 Speech synthesis techniques generate speech-like waveforms from textual words or symbols. Speech synthesis systems have been used for various applications, including speech-to-speech translation applications, where a spoken phrase is translated from a source language into one or more target languages. In a speech-to-speech translation application, a speech recognition system translates the acoustic signal into a computer- readable format, and the speech synthesis system reproduces the spoken phrase
15 in the desired language.

 FIG. 1 is a schematic block diagram illustrating a typical conventional speech synthesis system 100. As shown in FIG. 1, the speech synthesis system 100 includes a text analyzer 110 and a speech generator 120. The text analyzer 110 analyzes input text and generates a symbolic representation 115 containing linguistic information
20 required by the speech generator 120, such as phonemes, word pronunciations, phrase boundaries, relative word emphasis, and pitch patterns. The speech generator 120 produces the speech waveform 130. For a general discussion of speech synthesis principles, see, for example, S. R. Hertz, "The Technology of Text-to-Speech," Speech Technology, 18-21 (April/May, 1997), incorporated by reference herein.

25 In a concatenative speech synthesis system, stored segments of human speech are typically pieced together to produce the speech output. When an utterance is synthesized by the speech generator 120, the corresponding speech segments are retrieved, concatenated, and modified to reflect prosodic properties of the utterance, such

as intonation and duration. Each of the concatenated speech segments has an inherent natural pitch contour that was uttered by the speaker. However, when small portions of natural speech arising from different utterances in the segment database are concatenated, the resulting synthetic speech does not have a natural sounding pitch contour.

5 To produce natural-sounding speech, the speech generator 120 must produce acoustic values, durations, and pitch patterns that simulate properties of human speech. The acoustic values and durations of a speech segment depend on the neighboring segments, degree of syllable stress and position in the syllable. Pitch patterns are a function of linguistic properties of the utterance as a whole. Prediction of
10 the pitch patterns is an important aspect of generating natural-sounding speech.

Typically, the pitch contour of the concatenated segments are modified using a predefined pitch contour, using either a statistical or rule-based method, that is imposed on the synthetic speech using digital signal processing techniques. The desired contour is typically specified as one or more values per vowel or syllable. Thereafter, the
15 pitch contour values associated with each syllable are connected, for example, using a piece wise linear function, resulting in a continuous function of pitch versus time throughout the synthetic utterance.

While speech synthesis systems employing such pitch contour techniques perform effectively for a number of applications, they suffers from a number of
20 limitations, which if overcome, could greatly expand the performance and utility of such speech synthesis systems. Specifically, currently available speech synthesis systems 100 fail to produce speech that approaches a natural-sounding human. A need therefore exists for a speech synthesis system that utilizes a pitch contour resulting in a more natural-sounding speech.

25

Summary of the Invention

Generally, the present invention provides a speech synthesis system that utilizes a pitch contour resulting in a more natural-sounding speech. The present invention modifies the predicted pitch, $b(t)$, for synthesized speech using a low frequency energy booster. The low frequency energy booster interpolates the discrete pitch values, if necessary, and increase the amount of energy of the pitch contour associated with low frequency values, such as all frequency values below 10 Hertz. The amount of energy of the pitch contour associated with low frequency values can be increased, for example, by adding band-limited noise (a carrier signal) to the pitch contour, $b(t)$, or by filtering the pitch values with an impulse response filter having a pole at the desired low frequency value. The present invention serves to add vibrato to the original pitch contour, $b(t)$, and improves the naturalness of the synthetic waveform.

A more complete understanding of the present invention, as well as further features and advantages of the present invention, will be obtained by reference to the following detailed description and drawings.

Brief Description of the Drawings

FIG. 1 is a schematic block diagram of a conventional speech synthesis system;

FIG. 2 is a schematic block diagram of a speech synthesis system in accordance with the present invention;

FIG. 3 is a frequency spectrum illustrating a certain amount of bravado that is added to the original pitch contour, $b(t)$, in accordance with the present invention; and

FIG. 4 is a flow chart describing an exemplary concatenative text-to-speech synthesis system incorporating features of the present invention.

Detailed Description of Preferred Embodiments

FIG. 2 is a schematic block diagram illustrating a speech synthesis system 200 in accordance with the present invention. The present invention is directed to a method and apparatus for synthesizing speech that utilizes an improved pitch contour resulting in a more natural-sounding speech.

As shown in FIG. 2, the speech synthesis system 200 includes the conventional speech synthesis system 100, discussed above, as well as a low frequency energy booster 220. The conventional speech synthesis system 100 may be embodied as the ETI-Eloquence 5.0, commercially available from Eloquent Technology, Inc. of Ithaca, NY, as modified herein to provide the features and functions of the present invention. As shown in FIG. 2, the conventional speech synthesis system 100 includes a pitch predictor 210 that predicts the pitch, $b(t)$, of the utterance associated with the input text, in a known manner. As previously indicated, the predicted pitch, $b(t)$, provides a pitch value specified for each syllable.

According to a feature of the present invention, the predicted pitch, $b(t)$, is modified by the low frequency energy booster 220 to interpolate the discrete pitch values and increase the amount of energy of the pitch contour associated with low frequency values, such as below 10 Hertz. The amount of energy of the pitch contour associated with low frequency values can be increased, for example, by adding band-limited noise (a carrier signal) to the pitch contour, $b(t)$. In this manner, the use of the carrier signal contributes vibrato 310 to the original pitch contour, $b(t)$, as shown in FIG. 3, and improves the naturalness of the synthetic waveform.

Thus, in one implementation, the vibrato 310 corresponds to a periodic carrier waveform, $p(t)$, added to the pitch contour, $b(t)$. Thus, the pitch frequency, $f(t)$, of the speech 230 generated by the speech synthesis system 200 can be expressed as follows:

$$f(t) = b(t) + p(t),$$

where $p(t) = a \sin(\omega t + \Phi)$;

a = amplitude of the pitch variation;

$\omega = 2\pi f_r$; and

f_r = rate of pitch variation.

Thus, the pitch frequency, $f(t)$, corresponds to a narrow band, low
5 frequency noise signal. In one illustrative embodiment, the narrow band results in a
single low frequency sine wave; having a frequency, f_r , of 2.7 Hertz (Hz) and an
amplitude, a , of 10 Hz. Thus, the original pitch contour, $b(t)$, is varied by +/- 10 Hz at a
rate of 2.7 Hz. It is noted that these parameters may vary depending on the sex, dialect
and other speech parameters of the speaker associated with the synthesized speech. The
10 pitch frequency, $f(t)$, of the speech 230 generated by the speech synthesis system 200 can
be also expressed as the sum of its sinusoidal components.

FIG. 4 is a flow chart describing an exemplary implementation of a
concatenative text-to-speech synthesis system 400 incorporating features of the present
invention. As shown in FIG. 4, the user initially specifies the text he or she wishes to be
15 synthesized during step 410. The text specified by the user is then used during step 420
to select the segments of speech that will be concatenated during step 430 to form the
synthetic waveform.

The user-specified text is also used during step 450 to calculate the desired
pitch value for each syllable in the utterance using statistical methods. From the desired
20 pitch values a piece wise linear contour is formed during step 460, yielding the pitch
contour, $b(t)$, a function of pitch versus time. Each of the steps performed in obtaining
the pitch contour, $b(t)$, may be performed in a conventional manner, such as using the
techniques employed by the ETI-Eloquence 5.0, referenced above.

During step 470, a narrow band, low frequency noise signal, $p(t)$, is added
25 to the pitch contour, $b(t)$, obtained in the previous step, in accordance with the present
invention. The output of the summation of step 470 becomes the final pitch contour of
the synthesized waveform. Thereafter, the pitch of the concatenated segments is adjusted

during step 480 to exhibit the final contour. After the pitch has been adjusted, the synthetic speech is available to be sent to a file or speaker.

The present invention can manipulate the pitch contour, $b(t)$, in various ways to increase the amount of energy with low frequency components, such as below 10 Hz, as would be apparent to a person of ordinary skill in the art. In a further variation, the discrete pitch values associated with each syllable can be interpolated in accordance with a procedure that likewise increases the amount of energy with low frequency components. For example, the present invention can be accomplished by passing the pitch values through an appropriate filter to increase the low frequency energy, such as an impulse response filter having a pole at the desired f_r .

It is to be understood that the embodiments and variations shown and described herein are merely illustrative of the principles of this invention and that various modifications may be implemented by those skilled in the art without departing from the scope and spirit of the invention.

For example, we have mentioned the use of this invention in a concatenative speech synthesis system. However, any method of producing synthetic speech, for example, formant synthesis or phrase splicing, could also make use of the invention by including a method for predicting pitch at the syllable level and imbedding that contour in a narrow band, low frequency noise signal, as would be apparent to a person of ordinary skill in the art.